Asters
Conservatoire d'espaces naturels Haute-Savoie

Laboratoire EDYTEM
Environnements, Dynamiques et Territoires de la Montagne

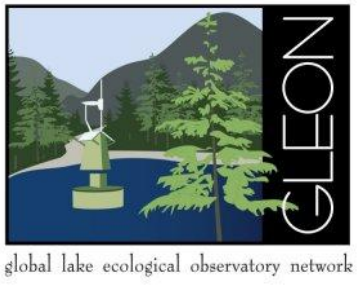
Avril 2014

Rosalie Bruel

Guide de l'utilisateur

Data Standardizer

Standardisation des données capteur



Programme libre développé à l'Université de Waikato (Nouvelle-Zélande) par Chris Mc Bride, Richard Lamont et Sam Shute. Il répond à la demande du GLEON concernant la nécessité de standardiser les données environnementales acquises dans le cadre de monitoring haute fréquence.

TABLE DES MATIERES

1. INTRODUCTION 1

2. IMPORT DU FICHIER 1

3. STANDARDISATION 2

4. EXPORT 3

5. FICHIERS D'EXPORTS 3

5.1. FICHIER DE DONNEES 3

5.1.1. Nom du fichier 3

5.1.2. Colonne(s) date et heure 3

5.1.3. Colonne(s) variable 4

5.2. FICHIER DE METADONNEES 4

5.2.1. Nom du fichier 4

5.2.2. Contenu du fichier 4

6. PROBLEMES EVENTUELLEMENT RENCONTRES… 5

6.1. SI LES UNITES SONT INDIQUEES POUR CHAQUE MESURE 5

6.2. FORMAT DATE/HEURE 6

6.3. AJOUTER UN NOUVEAU PARAMETRE 6

7. REFERENCES 6

8. ANNEXES 7

8.1. ABREVIATION DES PARAMETRES, UNITES RECOMMANDEES ET ALTERNATIVES 7

8.2. CODES NON DISPONIBLES A LA SELECTION (INVENTES POUR LES DONNEES DE MONITORING DU LAC D'ANTERNE) 8

1. Introduction

Data Standardizer est un exécutable (c'est-à-dire un fichier, donc une suite de données ou octets placée sur le disque dur comme n'importe quel fichier texte ou autre), qui permet de standardiser des jeux de données. Il est particulièrement utile pour standardiser les données haute-fréquence issues de programmes de monitoring (rivières, lacs, océans). Il peut éventuellement être utilisé pour « nettoyer » un fichier issus de profil de sonde, tout en prenant en compte qu'en standardisant les paramètres de date et heure la résolution de sortie est de 1 minute.

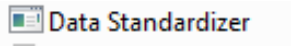
Il a été développé par Chris McBride, Richard Lamont et Sam Shute à l'université de Waikato (Nouvelle-Zélande). Il est particulièrement intéressant à utiliser avant de compiler des données de plusieurs capteurs/différentes années sur B3 (voir Manuel d'utilisation). Il utilise notamment le format de représentation standard ISO 8601 pour la date et l'heure. Il permet également de standardiser les noms des paramètres, et dispose d'un outil d'agrégation des données.

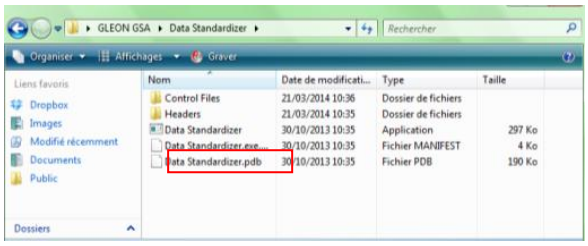
2. Import du fichier

Le format d'entrée est un fichier .txt (Bloc-notes) ou .csv.

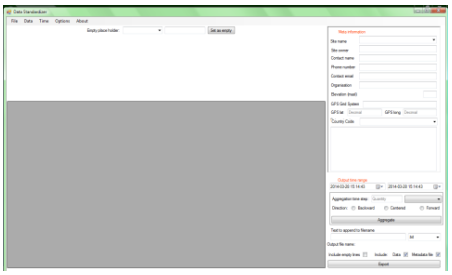
REMARQUE : Attention à conserver les chiffres significatifs (les décimales) si les données sont visualisées au préalable sur MS Excel : celui-ci n'affiche pas nécessairement toutes les décimales.

La marche à suivre est la suivante :

2.1. Ouvrir Data Standardizer 



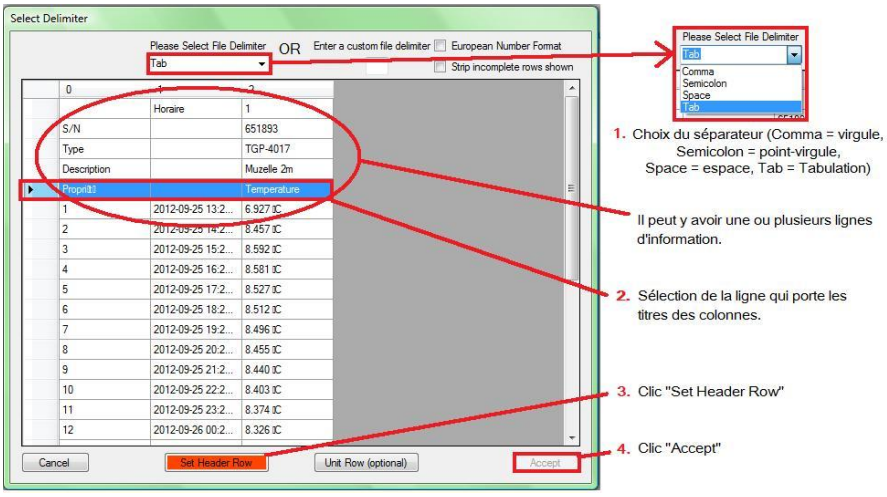
➔



2.2. File > Open data set

2.3. Charger le jeu de données brut en cliquant sur Open data.

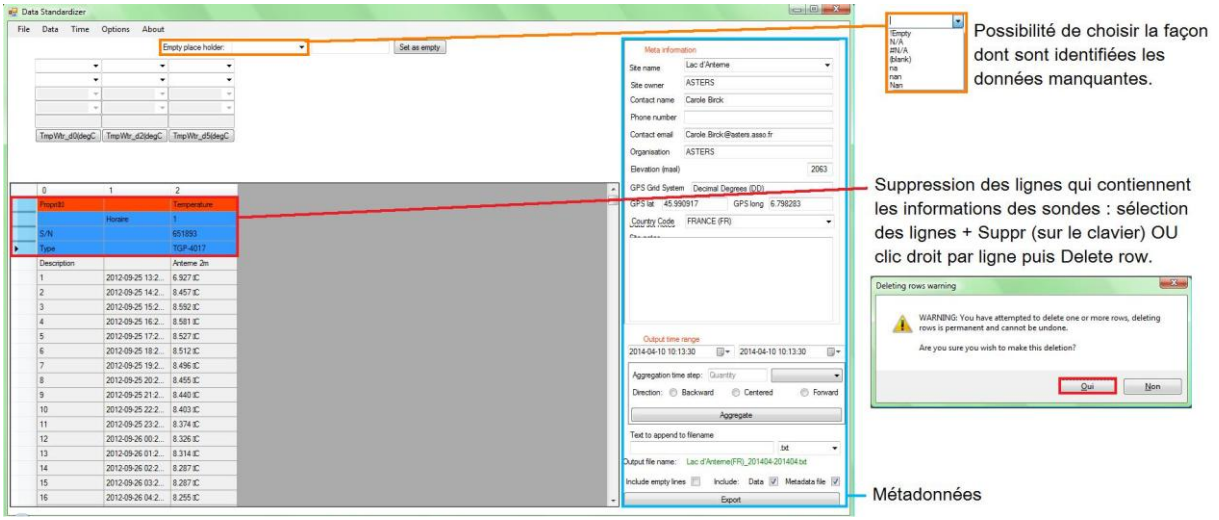
2.4. Sélection du séparateur.



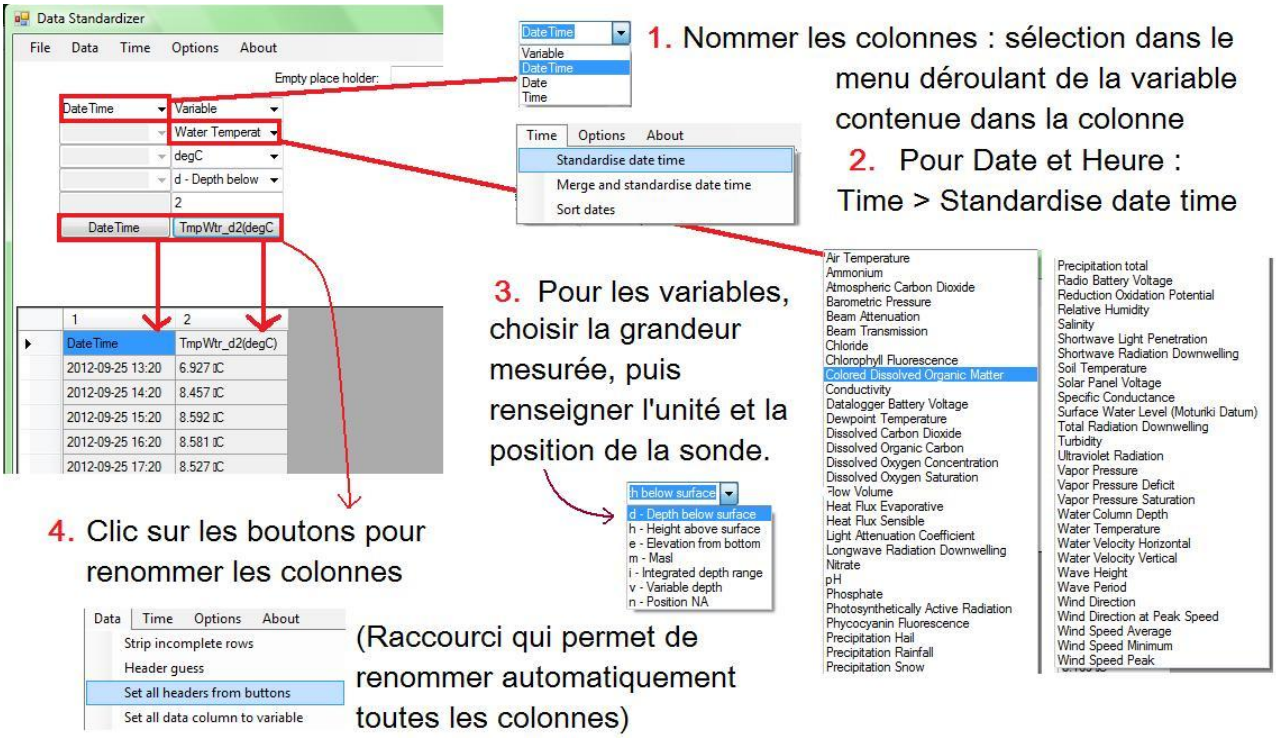
2.5. Si une fenêtre « No Data column was found. » s'ouvre, cliquer sur "Ok".

3. Standardisation

3.1. Supprimer les colonnes et les lignes inutiles (dans cet exemple, la première colonne peut être supprimée puisqu'elle indique seulement le numéro d'enregistrement).


Possibilité de choisir la façon dont sont identifiées les données manquantes.
Suppression des lignes qui contiennent les informations des sondes : sélection des lignes + Suppr (sur le clavier) OU clic droit par ligne puis Delete row.
Métadonnées

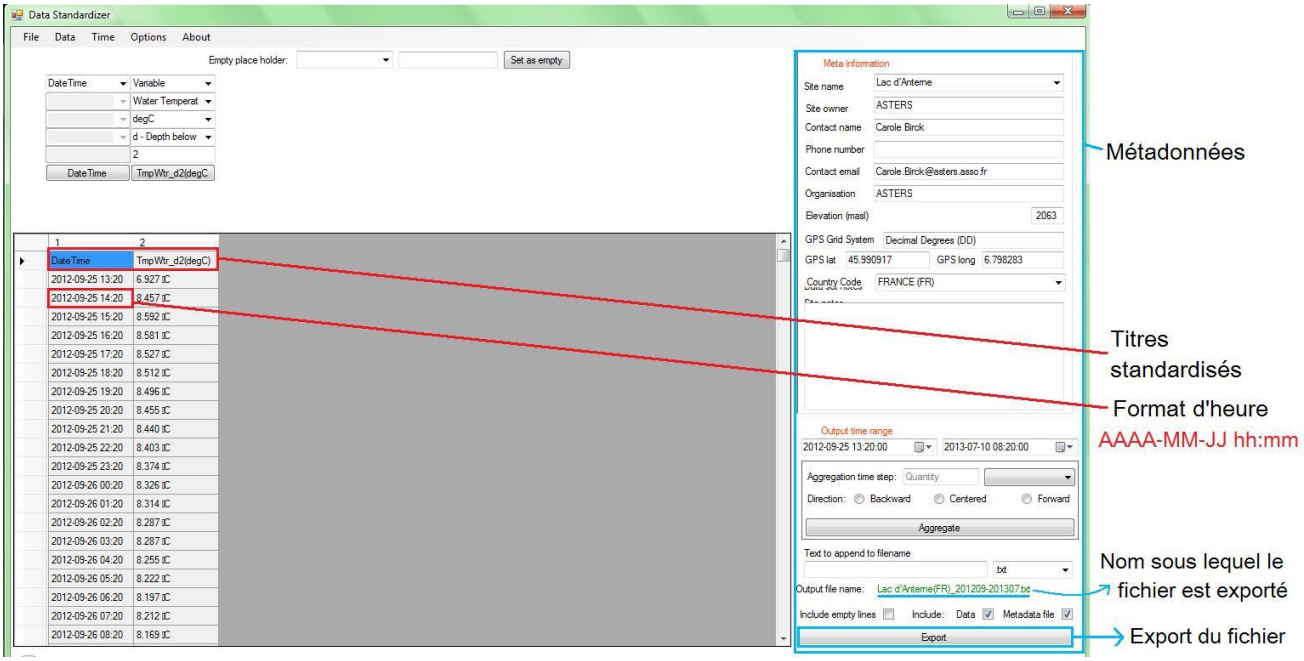
3.2. Standardiser les données :


1. Nommer les colonnes : sélection dans le menu déroulant de la variable contenue dans la colonne
2. Pour Date et Heure : Time > Standardise date time
3. Pour les variables, choisir la grandeur mesurée, puis renseigner l'unité et la position de la sonde.
4. Clic sur les boutons pour renommer les colonnes
(Raccourci qui permet de renommer automatiquement toutes les colonnes)

REMARQUE : Lorsqu'une unité n'est pas disponible, il est possible de double-cliquer sur la case contenant le nom du paramètre puis de le remplacer manuellement. C'est la procédure que j'ai suivi à chaque fois qu'un paramètre n'était pas disponible à la sélection. J'ai alors inventé un code pour ce paramètre tout en respectant la case choisie par les programmateurs de Data Standardizer (voir le paragraphe 5.1.3 qui explique la façon dont sont codés les paramètres). Il faut à ce moment veiller à bien utiliser le même code d'une année à l'autre. Les codes que j'ai utilisés pour les paramètres sont donnés en Annexe 2 (8.2).

4. Export

4.1. Le fichier ainsi standardisé peut être exporté en cliquant sur le bouton en bas à droite « Export ».



4.2. Un dossier de fichier est créé à l'emplacement choisi, celui-ci contient le fichier de données et un fichier de métadonnées.

4.3. Pour tout nouveau fichier, recommencer la procédure. Possibilité de charger un dossier de métadonnées renseigné à l'avance à l'étape 2.3. Cela rempli automatiquement les données du site ainsi que les titres.

5. Fichiers d'exports

Lors de l'export, un dossier se crée à l'endroit choisi (par exemple le bureau). Ce dossier comprend le fichier des données et celui des métadonnées, tous deux au format texte. Le délimiteur de colonnes est la tabulation.

5.1. Fichier de données

5.1.1. Nom du fichier

Le format de nom des fichiers de sortie est « Nom(XX)_AAAAMM-AAAAMM.xxx ».

Nom	Lac ou nom du site.
(XX)	Code du pays en deux lettres (France = FR).
AAAAMM-AAAAMM	Plage de temps des données.
Xxx	Extension du fichier (normalement .txt).

*Exemple :*Lac d'Anterne(FR)_201301-201312.txt comprend des données du Lac d'Anterne, en France, de janvier à décembre 2013.

5.1.2. Colonne(s) date et heure

La première colonne du fichier de données comprend la date et l'heure. Le titre de la colonne est Date Time. Le format est AAAA-MM-JJ hh:mm (*exemple :* 2014-04-16 09:45 pour une mesure enregistrée le 16 avril 2014 à 9h45).

Il est également possible de ne pas fusionner les colonnes de date et heure. Les formats date AAAA-MM-JJ et heure hh:mm restent les même (*exemple :* 2014-04-16 | 09:45). Les titres des colonnes sont Date et Time.

5.1.3. Colonne(s) variable

Le format des titres de colonnes est « XXXXXX_Lpp(UUU) ».

XXXXXX Deux à six caractères qui correspondent à l'abréviation du paramètre (voir Annexe 1 (8.1) pour la liste fournie par les développeurs de Data Standardizer.

L Indicateur de la position de la sonde :

- d = profondeur sous la surface en mètres.
- h = hauteur au-dessus de la surface en mètres.
- e = élévation à partir du fond du lac (pour les bouées) en mètres.
- m = altitude par rapport au niveau de la mer en mètres.
- i = position d'un dispositif d'échantillonnage de type tube en mètres (remplace « pp » par « pp-pp » pour indiquer la gamme de profondeur).
- v = profondeur variable, si un appareil la mesure. La colonne peut prendre le nom « Depth_v(m) ». Pour des dispositifs qui relèvent le profil du lac, pp n'est pas requis.
- n = la position de la sonde n'est pas une information applicable à la situation, ou n'est pas relevante (*exemple :* les variables dérivées comme la Stabilité de Schmidt).

Pp Profondeur, hauteur, élévation au-dessus du fond du lac, altitude de la sonde en mètres.

(UUU) Unité des mesures.

*Exemple :*TmpWtr_d2(degC) est la température de l'eau en degré Celsius à 2 mètres en dessous de la surface.

L'usage des points et des virgules comme séparateurs décimaux semble être accepté par Data Standardizer. Bien que le manuel rédigé par les créateurs du programme recommande l'utilisation des points, nous avons remarqué que B3 par exemple ne reconnait que les virgules (en tout cas dans la version de novembre 2013). Il est à noter que certains logiciels, comme R qui permet l'analyse de données, ne reconnaissent que les points.

Il est possible de choisir l'indicateur de données manquantes parmi !EMPTY, NaN, na, nan, #N/A, N/A.

Exemple : 3051.42 #N/A 23.56 432.9

5.2.Fichier de Métadonnées

5.2.1. Nom du fichier

Le format de nom des fichiers de sortie est « Nom(XX)_metadata.txt ».

Nom Lac ou nom du site.

(XX) Code du pays en deux lettres (France = FR).

5.2.2. Contenu du fichier

Plusieurs champs peuvent être complétés afin de renseigner au mieux le site.

Associated data file(s) Fichiers de données associés à ces métadonnées

Site Name Nom du site complet (*exemple :* Station météo Lac d'Anterne)

Owner Personne ou organisation qui détient les droits sur les données. Organisation qui a collecté les données.

Latitude/Northing *Exemple :* -39.46

Longitude/Easting *Exemple :* -62.701733

GPS Grid System Le système par défaut est WGS84 en degrés décimaux (facile à visualiser sur Google Earth).

Elevation (masl) Altitude du site de monitoring par rapport au niveau de la mer. « masl » est l'abréviation de meter above sea level.

Country A sélectionner à partir du menu déroulant.

Informations du contact pour ce site…

Name Nom du contact en cas de besoin de plus d'informations sur le site, le jeu de données, et la permission d'utilisation.
Organisation Structure qui emploie le contact.
Phone Numéro de téléphone du contact.
Email Email du contact.

Paramètres mesurés et informations complémentaires…

Number of Sensors Nombre de variables (sans compter la(les) colonne(s) date et heure qui sont comprises dans le jeu de données.

Variable names Nom des différentes variables du fichier de données.
Exemple :
TmpAir_h1.5(degC)
RelH_h1.5(%)
Etc.

Data set Notes Texte libre permettant la description du jeu de données. Ici peuvent figurer :
- La permission d'utilisation,
- Des informations pour l'assurance et le contrôle de qualité,
- Des personnes contacts additionnelles,
- Publications relatives aux données
- N'importe quelle autre information sur les données

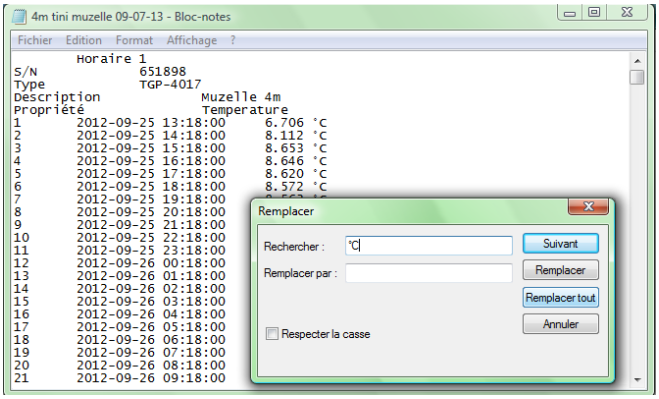
Site Notes Texte libre pour la description du site de suivi. Ici peuvent figurer :
- Plus d'informations sur la position et le fonctionnement des stations de mesures,
- Plus de détails sur les sondes,
- N'importe quelle autre information sur le site.

6. Problèmes éventuellement rencontrés…

Quelques problèmes ont été rencontrés lors de la prise en main de Data Standardizer, et puisqu'aucun document officiel n'existe pour le moment, voici un résumé de nos observations.

6.1. Si les unités sont indiquées pour chaque mesure

Les cases comportant les mesures des sondes ne doivent afficher que des nombres, et non pas les unités. Les fichiers d'export par défaut de Tinytag Explorer comportent par exemple l'unité, à moins de le préciser lors de l'enregistrement (dans Tinytag Explorer : enregistrer sous > Fichier texte séparé par une tabulation (sans unité) (*.txt)). Il est également possible de supprimer les unités dans Bloc-notes (Barre de menu : Edition > Remplacer > Dans la case « Rechercher » : mettre l'unité ; dans la case « Remplacer par » : ne rien mettre > Remplacer tout).

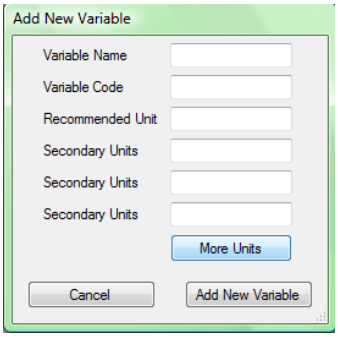


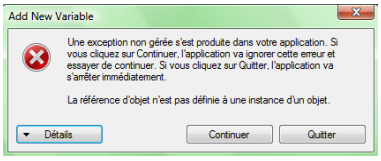
6.2.Format date/heure

Attention au format des dates lors de l'import sur Data Standardizer : le logiciel reconnait un grand nombre de format, mais ne reconnaitra pas de date si le mois est écrit en lettre (par exemple 10 juil. 2013 au lieu de 10/07/2013). Parfois MS Excel va également remplacer un format de date, interpréter « 05/07/08 » par 8 juillet 2005 au lieu de 5 juillet 2008. Attention donc si vous avez d'abord ouvert votre fichier sur Excel, qui a tendance à modifier la date. Cependant les fichiers de sortie de sonde sont généralement bien formatés.

6.3.Ajouter un nouveau paramètre

Data Standardizer propose d'ajouter de nouveaux paramètres ou de nouvelles unités (Barre de menu : Options > Add New Parameter). Il faut alors renseigner la description du paramètre, le code choisi, l'unité recommandé (Système International), et éventuellement d'autres unités. Cependant un message d'erreur apparait, il est donc impossible de finir la procédure.




Add New Variable
Une exception non gérée s'est produite dans votre application. Si vous cliquez sur Continuer, l'application va ignorer cette erreur et essayer de continuer. Si vous cliquez sur Quitter, l'application va s'arrêter immédiatement.
La référence d'objet n'est pas définie à une instance d'un objet.
Détails
Continuer
Quitter

Je n'ai pas trouvé de solution à ce jour. Il faut donc ajouter manuellement les paramètres en double-cliquant sur la case de titre comme expliqué dans la remarque du point 3.2.

7. Références

McBride C. (2013). Draft – GLEON – Data Standardizer. Document non publié, obtenu par Marie-Elodie Perga (INRA) lors de la conférence GLEON organisée en novembre 2013. 3p

8. Annexes

8.1.Abréviation des paramètres, unités recommandées et alternatives

GLEON code	GLEON control vocab	Recommended units	Alternative units
BaroP	Barometric_Pressure	hPa	mmHg, mbar, kPa
BFrq	Bouyancy_Frequency		
CDOM	Colored_Dissolved_Organic Matter	ug/L	
CDOMCR	Colored_Dissolved_Organic Matter_Coefficient_of_Range		
Chl	Chlorophyll a	ug/L	
ChlF	Chlorophyll_Fluorescence	RFU	RFUB, ug/L chl
Cl	Chloride	mg/m^3	g/m^3,ppm,ppb
CO2A	Atmospheric_Carbon_Dioxide	ppm	
Cond	Conductivity	mS/cm	uS/cm,
CondSp	Specific_Conductance	mS/cm	uS/cm
Depth	Depth	m	ft
DIN	Dissolved_Inorganic_Nitrogen		
DCO2	Dissolved_Carbon_Dioxide	mg/L	ppm
DO	Dissolved_Oxygen_Concentration	mg/L	ppm
DO	Dissolved_Oxygen_Saturation	%sat	
DOC	Dissolved_Organic_Carbon	mg/L	ppm
Hail	Precipitation_Hail	hits/cm^2	
HFlxEv	Heat_Flux_Evaporative	W/m^2	
HFlxSn	Heat_Flux_Sensible	W/m^2	
Kd	Light_Attenuation_Coefficient	m^-1	
NH4	Ammonium	mg/m^3	g/m^3,ppm,ppb
NNN	Nitrate_+_Nitrite	mg/m^3	g/m^3,ppm,ppb
NO2	Nitrite	mg/m^3	g/m^3,ppm,ppb
NO3	Nitrate	mg/m^3	g/m^3,ppm,ppb
ORP	Reduction_Oxidation_Potential	mV	
PARP	Photosynthetically_Active_Radiation_Penetration		
pH	pH		
PhyF	Phycocyanin_Fluorescence	RFU	RFUB, cells/mL
PO4	Phosphate	mg/m^3	g/m^3,ppm,ppb
Ppn	Precipitation	mm	
RadLWD	Longwave_Radiation_Downwelling	W/m^2	
RadPAR	Photosynthetically_Active_Radiation	umol/m^2/s	
RadSWD	Shortwave_Radiation_Downwelling	W/m^2	
RadSWP	Shortwave_Light_Penetration		
RadTD	Total_Radiation_Downwelling	W/m^2	
RadUV	Ultraviolet_Radiation	mW/cm^2	
Rain	Precipitation_Rainfall	mm	m
RelH	Relative_Humidity	%	
Sal	Salinity	PSU	TDS
Snow	Precipitation_Snow	mm	
SoilT	Soil_Temperature	degC	degF
TKN	Total_Kjeldahl _Nitrogen	mg/m^3	g/m^3,ppm,ppb
TmpAir	Air_Temperature	degC	degF
TmpDew	Dewpoint_Temperature	degC	degF
TmpLg	Datalogger_Temperature	degC	degF
TmpRef	Reference_Temperature	degC	degF
TmpWtr	Water_Temperature	degC	degF
TN	Total_Nitrogen	mg/m^3	g/m^3,ppm,ppb
TOC	Total_Organic_Carbon	mg/L	ppm
TP	Total_Phosphorus	mg/m^3	g/m^3,ppm,ppb
Tran	Light_Transmittance	%	
TranCR	Transmittance_Coefficient_of_Range		
TSS	Total_Suspended_Solids	g/m^3	
Turb	Turbidity	FTU	RTU, NTU
VBatLg	Datalogger_Battery_Voltage	V	
VBatR	Radio_Battery_Voltage	V	
VBatS	Sensor_Battery_Voltage	V	
VP	Vapor_Pressure	hPa	mbar, kPa
VPDef	Vapor_Pressure_Deficit	hPa	mbar, kPa
VPSat	Vapor_Pressure_Saturation	hPa	mbar, kPa
VSol	Solar_Panel_Voltage	V	
WaveHt	Wave_Height	m	ft
WavePd	Wave_Period	s	
WndDir	Wind_Direction	deg	
WndDrP	Wind_Direction_at_Peak_Speed	deg	
WndGst	Wind_Peak_Speed	deg	
WndSpd	Wind_Speed	m/s	kt, km/hr
WtrLvl	Water_Level	m	ft
WtrVlH	Water_Velocity_Horizontal	m/s	cm/s
WtrVlV	Water_Velocity_Vertical	m/s	cm/s

8.2. Codes non disponibles à la sélection (inventés pour les données de monitoring du lac d'Anterne)

Les codes suivants respectent au maximum le Thesaurus utilisé par Chris McBride *et al* lors de la programmation de Data Standardizer. Les unités sont celles issues des capteurs, sans conversion.

Station	Description	Abbréviation	Unités	Position	Détail
Decagon	Soil Moisture	SoilM_bareN(m3/m3) / SoilM_bareS(m3/m3)	m^3/m^3	Position (Bare : surface nue, Vegetated : surface végétalisée) Orientation Nord ou Sud	Bare (Nord ou Sud)
	Soil Moisture	SoilM_vegN(m3/m3) / SoilM_vegS(m3/m3)	m^3/m^3		Vegetated (Nord ou Sud)
	Soil Temperature	SoilT_bareN(degC) / SoilT_bareS(degC)	°C		Bare (Nord ou Sud)
	Soil Temperature	SoilT_vegN(degC) / SoilT_vegS(degC)	°C		Vegetated(Nord ou Sud)
Hydro	Water Height	WatLvl_n(cm)	cm	Position sans importance : code n	
	High turbidity	TurbHi_n(mv) / TurbHi_n(FTU)	mV / FTU		
	Low turbidity	TurbLw_n(mv) / TurbLw_n(FTU)	mV / FTU		
RBR	Water pressure	WtrP_d12(dbar)	Decibar	d - Depth below surface	12
	Chlorophyll Fluorescence	FlChl_d12(mcg)	µg/l		
	Depth	Depth_v(m)	m		
Météo	Solar radiation	SolarR_n(kW/m2)	W/m^{-2}	Position sans importance : code n	
	Standard deviation of the wind direction	SDWnd_n(deg)	deg		
	Snow Height	Snow_n(cm)	cm		
	Air Pressure	BaroP_n(mV)	mV		
	Air Pressure	BaroP_n(mbar)	mbar		